Figure 1. The model legume *Medicago truncatula*.
The shape of *M. truncatula* seed pods (top right) lead to its common name of barrel medic.

genetics, *M. truncatula* harbors several attributes that make it attractive as a model species. It has a short seed-to-seed generation time and abundant seed set. Extensive collections of *M. truncatula* ecotypes — natural geographic variants — exist, and mutant collections are readily produced by physical (fast neutron) and transposon-mediated mutagenesis. *M. truncatula* is also host to *Sinorhizobium meliloti*, a rhizobium species whose genome has been fully sequenced.

What resources are available for* M. truncatula *research? With almost 200,000 expressed sequence tags (ESTs) and greater than 85 Mb of genome sequence currently deposited in publicly available databases, legume researchers have access to a rich data set that outlines the gene content of *M. truncatula*. In addition, community-standardized *M. truncatula* DNA microarrays are available which facilitate near global transcript profiling. Many genes in *M. truncatula* have over 98% sequence identity with their orthologs in alfalfa, and so the *M. truncatula* microarrays can also be used for expression profiling of alfalfa. Collaborative research programs are underway that detail the transcript, protein and metabolite profiles of *M. truncatula*. A series of forward and reverse genetics methodologies and populations have also been established. Extensive bioinformatics resources also exist which facilitate data comparison within *M. truncatula* and leverage the data available for this model species to other agronomically important legume crop species.

Is there a* M. truncatula *genome project? After *Arabidopsis* and rice, *M. truncatula* will be the next plant to have its genome completely sequenced. An international *Medicago* sequencing project has begun that involves laboratories in the US, the UK and France. It is anticipated that sequencing the *Medicago* genome will be completed within the next three years. The genome sequence of *M. truncatula* will serve as a basis for structural genomics comparisons with other legume species such as alfalfa and soybean.

Where can I find out more about* M. truncatula*?

Cook, D.R. (1999). *Medicago truncatula* – a model in the making! Curr. Opin. Plant Biol. *2*, 301–304.
Young, N.D., Mudge, J., and Ellis, T.N. (2003). Legume genomes: more than peas in a pod. Curr. Opin. Plant Biol. *6*, 199–204.
Dixon, R.A., and Sumner, L.W. (2003). Legume natural products. Understanding and manipulating complex pathways for human and animal health. Plant Physiol. *131*, 878–885.
The Center for *Medicago* Genomics Research: www.noble.org/medicago/index.html
The Consensus Legume Database. www.legumes.org
Medicago genome project at the University of Oklahoma. www.genome.ou.edu/medicago.html
The Legume Information System. www.comparative-legumes.org
Toulouse, C.N.R.S.-I.N.R.A. http://medicago.toulouse.inra.fr/Mt/EST/
The Medicago truncatula Gene Index. www.tigr.org/tdb/tgi/mtgi/
Medicago Bioinformatics at the University of California – Davis. http://medicago.plantpath.ucdavis.edu/

Plant Biology Division, The Samuel Roberts Noble Foundation, 2510 Sam Noble, Parkway, Ardmore, Oklahoma 73401, USA. E-mail: gdmay@noble.org

Correspondence

‘Spalog’ and ‘sequelog’: neutral terms for spatial and sequence similarity

Alexander Varshavsky

Similarities amongst sequences or three-dimensional (3-D) structures and conjectures based on similarities are a major topic of molecular biology and related fields. Therefore it is striking that there are presently no terms that denote a sequence or a 3-D structure that is similar to another sequence or 3-D structure without implying *anything at all* about evolutionary relatedness or biological functions. The lack of such *neutral* terms for denoting similarity is one reason for the widespread use of the terms ‘homolog’, ‘ortholog’ and ‘paralog’. The first term is more than a century old and the other two were proposed long before the advent of extensive sequence comparisons [1].

To state that a gene or a protein *A* is a homolog of *B* implies that *A* and *B* are related through common descent, a proposition that needs to be proven in most cases [2]. In addition, two sequences can be 37% identical, but they cannot be 37% homologous — they are either homologous or not. The frequent unsuitability of the term ‘homolog’ in the context of similarity was pointed out repeatedly [2,3], but the literature is still rife with this misuse, in part because proper neutral terms simply do not exist.

The disposition can be also difficult with the terms ‘ortholog’ and ‘paralog’. Orthologs are two homologous sequences that diverged following *speciation*, such that the common precursor of two sequences was harboured by the last common ancestor of the two species. Paralogs, by contrast, are two homologous

sequences that diverged after *gene duplication* [1]. Besides the initial ambiguity in assigning homology — two orthologs are homologous, and two paralogs are homologous as well — the use of 'ortholog' and 'paralog' implies additional probabilistic inferences about the evolution of the two sequences being compared [4,5]. Yet further imprecisions often accrue, because throughout the literature the terms 'ortholog' and 'paralog' are also used to denote *functional similarities* between orthologous genes (e.g., similar enzymatic activities of their protein products) and *functional differences* between paralogous genes. Neither of these relationships, which are often presumed but not proven, is implied by the definitions of 'ortholog' and 'paralog'.

To say that the current usage of 'homologs', 'orthologs' and 'paralogs' is complicated and often less than rigorous would be understating the case. A statistically significant similarity is an experimental fact, whereas 'homology', 'orthology' and 'paralogy' are, in most cases, hypotheses. There is, at present, a striking disparity between the generally high rigor of statistical methods used to compare sequences or structures and the often cavalier, assumption-laden attitude in the use of 'homolog', 'ortholog' and 'paralog'.

To remedy this, I propose two terms to denote similarity, '*sequelog*' and '*spalog*'. They meet the requirement of evolutionary and functional neutrality, are mnemonically helpful, and make it possible to distinguish through single words between the realms of similar sequences and similar 3-D structures.

The term '*sequelog*' denotes a nucleotide or amino acid sequence that is similar, to a specified extent, to another sequence. The term '*spalog*' (pronounced [*spailog*]) denotes a 3-D structure that is spatially similar, to a specified extent, to another 3-D structure. These terms are strictly about similarity and imply nothing about evolutionary relatedness and functional properties of the sequences or structures.

In comparing nucleotide or amino acid sequences, the extent of similarity is conveyed by a numerical score, the percentage of nucleotide or amino acid positional identity. Alternatively, the extent of similarity of two sequences can be conveyed by the probability of an identical score for a randomly chosen pair of sequences of the same length. In comparing amino acid sequences, one can also measure the percentage of similarity (in contrast to the percentage of identity). This includes the identical residues as well as residues that are scored as 'similar' to corresponding residues of the second sequence, according to a similarity matrix [2,6].

When the 3-D structures of two proteins or nucleic acids are compared, a standard measure of similarity is the root-mean-square deviation (r.m.s.d.) between atomic positions. In principle, one could introduce the term 'similog' to denote either a sequence or a 3-D structure that is similar to another sequence or 3-D structure. However, the considerable advantage of 'sequelog' and 'spalog' is that they instantly define the nature of similarity (a sequence or a spatial one), thus obviating further clarifications.

In a typical usage of the proposed terms, one can state, for example, that protein *A* is a sequelog of protein *B* (*x*% identity over *y* residues), or that protein *C* is a spalog of protein *D* (r.m.s.d. of *x* Å for *y* equivalent C_α atoms). Related measures of spatial similarity include a Z-score computed with the program DALI [7]. To add qualitative, shorthand distinctions one can state, for example, that protein *A* is a weak but significant sequelog of protein *B* (e.g., 24% identity over 165 residues), or that protein *C* is a strong spalog of protein *D* (e.g., r.m.s.d. of 2.3 Å for 120 equivalent C_α atoms), or that protein *E* is a strong sequelog of protein *F* (e.g., 60% identity over 372 residues). In using the 'sequelog' terminology, it would be best to invoke just the percentage of identity of two sequences and its statistical significance, which is straightforwardly computable. This would avoid the influence of any other information, for example, similarity matrices or 3-D structures. Yet again, the central idea is to minimize 'interpretational' aspects of 'sequelog', 'spalog' and the derivative terms, such as 'sequelogy', 'sequelogous', 'spalogous' and so forth.

A strong sequelog of a given protein is very likely to be its spalog as well [8], but the converse is not necessarily true, as a strong spalog of a given protein may not be its sequelog. For example, the 66-residue *Escherichia coli* protein ThiS, the sulfur carrier in the pathway of thiamine biosynthesis, is a strong spalog of the 76-residue eukaryotic ubiquitin (r.m.s.d. of 2.4 Å over 63 equivalent C_α atoms and a high Z-score of 5.2). However, it is not a sequelog of ubiquitin, as the sequence similarity between the two proteins (14%) is statistically insignificant, without additional information from 3-D structures [9]. Such comparisons can also employ the adjectives 'sequelogous' or 'spalogous'. For example, 'spalogous' can be used to denote similar local 3-D folds in otherwise dissimilar proteins. Thus: 'Although protein *A* is not a sequelog of protein *B*, the 23-127 region of *A* is strongly spalogous to the 769-875 region of *B* (r.m.s.d. of 2.2 Å over 101 equivalent C_α atoms, and Z-score of 5.6)'.

To describe a comparison of sequences or 3-D structures, one can begin by using 'sequelog' and 'spalog' or their derivatives in stating and numerically specifying the facts of similarity, as described above. After that, and only after that, one can conjecture, if necessary, based on additional evidence, that the two sequelogs or spalogs are likely to be 'homologs', 'orthologs', 'paralogs', or whatever. This way, the rigorous, numerically explicit statements about similarities of specific sequences or 3-D structures won't be conjoined, at birth, with the often unproven inferences that the latter three terms inherently imply.

Spalog, *sequelog* and terms derived from them fill an overt lacuna in the existing terminology. These terms would clarify and streamline discourses about similarity.

References

1. Fitch, W. (1970). Distinguishing homologous from analogous proteins. Syst. Zool. *19*, 99–113.
2. Koonin, E.V., and Galperin, M.Y. (2002). Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. (Boston: Kluwer Academic Publishers).
3. Reeck, G.R., de Haën, C., Teller, D.C., Doolittle, R.F., Fitch, W.M., Dickerson, R.E., Chambon, P., McLachlan, A.D., Margoliash, E., Jukes, T.H., et al. (1987). Homology in proteins and nucleic acids: a terminology muddle and a way out of it. Cell *50*, 667.
4. Sonnhammer, E.L., and Koonin, E.V. (2002). Orthology, paralogy, and proposed classification for paralog subtypes. Trends Genet. *18*, 619–620.
5. Jensen, R.A. (2001). Orthologs and paralogs - we need to get it right. Genome Biol. *2*, 1002.
6. von Heijne, G. (1987). Sequence analysis in molecular biology (New York: Academic Press).
7. Holm, L., and Sander, C. (1995). Dali: a network tool for protein structure comparison. Trends Biochem. Sci. *20*, 478–480.
8. Abagyan, R.A., and Batalov, S. (1997). Do aligned sequences share the same fold? J. Mol. Biol. *273*, 355–368.
9. Wang, C.C., Xi, J., Begley, T.P., and Nicholson, L.K. (2001). Solution structure of ThiS and implications for the evolutionary roots of ubiquitin. Nat. Struct. Biol. *8*, 47–51.

Division of Biology, 147–75, California Institute of Technology, 1200 E. California Boulevard, Pasadena, California 91125, USA. Email: avarsh@caltech.edu

Correspondence

Evidence for a Hox14 paralog group in vertebrates

Thomas P. Powers and Chris T. Amemiya*

The genealogy of vertebrate Hox genes and clusters has long fascinated evolutionary and developmental biologists alike. The importance of Hox genes is underscored by their involvement in axial patterning, their spatial colinearity of gene order with respect to expression domains and their likely interconnection with morphological evolution [1]. Whereas the protovertebrate amphioxus possesses a group 14 gene at the 5′-end of its 'archetypal' Hox cluster (*AmphiHox14*) [2,3], such a group 14 gene has thus far not been identified in any vertebrate. Based on these observations, the ancestral condition for the jawed vertebrates (gnathostomes) has been inferred to consist of 13 paralogous groups of Hox genes. According to this scenario, the *Hox14* gene of amphioxus would have originated by tandem duplication of a posterior Hox gene in this lineage [3].

Sequence analysis of the HoxA cluster of the Indonesian coelacanth (*Latimeria menadoensis*) and the HoxD cluster of the horn shark (*Heterodontus francisci*) revealed in each case an additional Hox gene between the group 13 gene and the *even-skipped* (*Evx*) ortholog (Figure 1A). These additional genes have the same transcriptional orientation as the other Hox genes in the cluster, but the opposite orientation to the *Evx* gene. *Hoxa14* and *Hoxd14* encode predicted proteins of 232 and 266 amino acids, respectively (Figure 1B). Moreover, these Hox14 genes possess 'split' homeoboxes and show exon-intron boundaries at identical positions as dipteran *Abdominal-B* genes and two of the 'posterior' Hox genes in amphioxus, including *AmphiHox14* [3]. These positions are different from those in the *Evx* genes, which also possess split homeoboxes; this further confirms that the Hox14 genes are not duplicated *Evx* orthologs. In addition, a *Hoxa14* pseudogene was found upstream of the *Hoxa13* gene in the horn shark. The exon structure of this pseudogene is similar to that of the *Hoxa14* and *Hoxd14* genes of the coelacanth and the horn shark, respectively (Figure 1A, and supplemental data). No other Hox14 genes were identified in surveys of the GenBank database.

In order to test the relationship of Hox14 genes to other posterior Hox genes, we constructed phylogenetic trees using amino acid sequences of the homeodomains as well as the complete proteins. Figure 1C shows a phylogenetic tree based on the homeodomains of horn shark *Hoxd14* and coelacanth *Hoxa14* with those of vertebrate group 13 and amphioxus *Hox13* and *Hox14*. Regardless of the phylogenetic method, all trees yield similar topologies and show a strong relationship of the two vertebrate Hox14 sequences to one another, but not to other posterior Hox genes or to any amphioxus sequences (Figure 1C, and supplemental data). This latter point raises questions as to the orthology of the *AmphiHox14* and gnathostome Hox14 genes, despite similar genomic structure and similar location within the respective clusters. However, the substantial amount of time that has passed since the divergence of amphioxus and gnathostomes (> 600 mya [4]), and the accelerated rate of molecular evolution of 'posterior' Hox genes may obscure a meaningful phylogenetic signal between vertebrate and amphioxus genes, thereby rendering such analyses problematic [3].

The shared identity between shark *Hoxd14* and coelacanth *Hoxa14* is emphasized by the partial alignment in Figure 1D. The high degree of relatedness has been retained despite involving two separate Hox clusters (A and